Can Mass/Count syntax be derived from semantics?

Ritwik Kulkarni

SISSA, Cognitive Neuroscience, Trieste, Italy

ritwik.kulkarni@yahoo.com

Alessandro Treves

SISSA, Cognitive Neuroscience, Trieste, Italy

ale@sissa.it

Susan Rothstein[z.l.]

Bar Ilan University, Gonda Multidisciplinary Brain Research Center, Ramat Gan, Israel

[z.l.]we dedicate this contribution to the memory of our cherished colleague, Susan, who passed away in the summer of 2019.

ABSTRACT

Analysing aspects of how our brain processes language may provide, even before the language faculty is really understood, useful insights into higher order cognitive functions. We have takeninitial steps in this direction, focusing on the mass-count distinction. The mass-count distinction relates to the countability or un-countability of nouns in terms of their syntactic usage.Our first results show that the mass-count distinction, across a number of natural languages, is far from bimodal, andexhibits in fact complex fuzzy relations between syntax and semantics. We then tried to test the ability of a

standard, biologically plausible self-organising neural network to learn such associations between syntax and semantics. A neural network that expresses competition amongst output neurons with lateral inhibition can identify the basic classes of mass and count in the syntactic markers and produce a graded distribution of the nouns along the mass-count spectrum. The network however fails to successfully map the semantic classes of the nouns to their syntactic usage, thus corroborating the hypothesis that the syntactic usage of nouns in the mass-count domain is not simply predicted by the semantics of the noun.

Keywords

Mass-Count distinction, Syntax-semantics interaction, Self-organisation, Neural networks

1. Introduction

The question of how the brain acquires language can be posed in terms of its ability to discover, from exposure to a corpus, the syntactic structure of a specific natural language and its relation with semantics. This has been a subject of study and of intense debate for the past few decades(Pinker S, 1995). Natural language acquisition appears to presuppose certain cognitive abilities like rule recognition, generalisation and compositional processing. These high-level abstract capabilities should be realized in the language domain and in specific sub-domainsof the language domain by general-purpose neural processing machinery, since there is no evidence for dedicated circuitry of a distinct type for each sub-domain of linguistic competencenor, for that matter, for language in general. How can rule recognition and generalization be implemented in standard, non-domain specific neural networks? To explore this issue, we focus our attention on a particular area ofthe syntax/semantics interface, the mass-count distinction. Following up on our previous study, where we investigated statistical aspects of the mass-count distinction in 6 languages, with relation to its cross-linguistic syntactic and semantic properties, we now aim to study the learnability of those syntactic properties by a basic neural network model, with the long-term goal of eventually understanding how such

processes might be implemented in the brain.The intuitively plausible first assumption is that mass nouns denote substance or ‘stuff' and do not denote individuated objects, whereas count nouns denote atomic entities that can be easily individuated and counted (Soja et al 1991, Pelletier 2010, Bale & Barner 2009, Chierchia 2010, Prasada et al 2002). This semantic difference seems to be reflected in the syntactic usage of the nouns in many natural languages since mass nouns and count nouns are associated with a different array of syntactic properties which are plausibly connected with countability and individuation. For example, in English, mass nouns are associated with quantifiers like *some, much* as in *some flour, much water,* cannot be directly modified by numerals *(*three flour(s))* and require a measure classifier (*kilos, bottles*) when used with numerals as in *six kilos of flour, three bottles/litres of water.* On the other hand count nouns are associated with determiners like*a/an,* (*a boy, an owl*), quantifiers like *many/few* (*many books, few people*) and crucially can be used with numerals without a measure classifier*three books, four boys.*. The traditional approach to the semantics of the mass-count distinction is that it can be expressed through properties of atomicity, cumulativity and homogeneity (Link 1983) .Count nouns are said to be atomic. A noun is atomic when its denotation includes distinguishable smallest elements which cannot be further divided into objects which are also in the noun denotation. So *chair* is count since it includes minimal chair-objects in its denotation which cannot be subdivided into smaller chairs, and plural *chairs* inherits atomicity from its singular stem Mass nouns are said to be cumulative and homogeneous. A noun is cumulative if the sum of two separate entitiesin the noun denotation is still in the denotation of the singular noun. For example if A is water and B is water then A and B together are water. Singular count nouns are not cumulative,since if A is in *chair* and B is in *chair* the sum of A and B is not in the denotation of singular *chair.* A noun is homogeneous if an entity its denotation can be subdivided into parts which are also in its denotation. For example, a part of something which is water is water, while a part of an object in *chair* is not a chair. So mass nouns are non-atomic and exhibit properties of being homogeneous and cumulative, whereas count nouns have opposite properties. However, as many linguists have pointed out, a simplistic mapping between homogeneity and mass syntax and/or

atomicity and count syntax on the other would imply that the expressions in different languages denoting the same real world objects would be consistently count or mass cross-linguistically. This is not the case. As we showed in (Kulkarni et al 2013), words with a similar interpretation may be associated with very different arrays of syntactic properties cross-linguistically. A noun which is associated with a count array of syntactic properties in one language may not be associated with a count array in a different language. Furthermore, over a sample of 6 different languages we saw that there is no binary divide into mass/count nouns, but rather a continuum with a group of nouns which are count with respect to all relevant properties, and then a range of nouns which are more of less count depending on how many count properties they display. This places the mass-count distinction at an interesting interface between the semantic properties of nouns and the syntax, since it raises the question of (i) what semantic properties are associated with count and mass syntax respectively, (ii) why there is variation in the noun categorization as mass or count cross-linguistically and (iii) how the knowledge of what is mass and count in a particular language is acquired.

2. Statistical analysis of cross-linguistic distribution of mass and count nouns

2.1. Data collection and distance distribution from pure count nouns

In a previous study (Kulkarni et al, 2013) we collected a database of how 1,434 nouns are used with respect to the mass/count distinction in six languages; additional informants characterized the semantics of the underlying concepts. A set of yes/no questions was prepared, in each language, to probe the usage of the nouns in the mass/count domain(e.g. does it occur with numerals, does it pluralise etc.). The questions probed whether a noun from the list could be associated with a particular morphological or syntactic marker relevant in distinguishing mass/count properties. A similar set of questions probed the semantic usage of the nouns using questions regarding the semantics properties of the nouns relevant for the mass-count distinction. Thus each nouns was associated with a binary string of 1 (Yes) and 0 (No), indicating how that particular noun is used in the mass-count space by

an informant. Since the data thus obtained is high dimensional in principle, as a first approximation, we consider the hypothesis that most of the information is contained on a single dimension of 'mass' and 'count'. We collapse the high dimensional data onto a single dimension (named as the MC dimension) by calculating the Hamming distance, or fraction of discordant elements, of each noun (i.e. of each syntactic group) from a bit string representing a pure count noun. A pure count string is one which has ‘yes’ answers for all questions identifying so-called ‘count’ properties and ‘no’ answers for all questions probing ‘mass’ properties. By plotting the distribution of nouns on this dimension we could provide a visualization of the main mass/count structure, to relate it with a linguistic interpretation. Thus a high dimensional numerical data has been compressed, albeit with some loss of information, to a representation that can readily identify the degree of “countness” of a noun. The resulting distribution of mass/count syntactic properties is seen to be graded in nature instead of either a binary or a bimodal distribution, as one might have expected intuitively. Most common nouns are strictly count in nature, in five of the six languages considered, with mass features increasingly rarer as they approach the pure mass ideal (See Kulkarni et al 2013 for details)

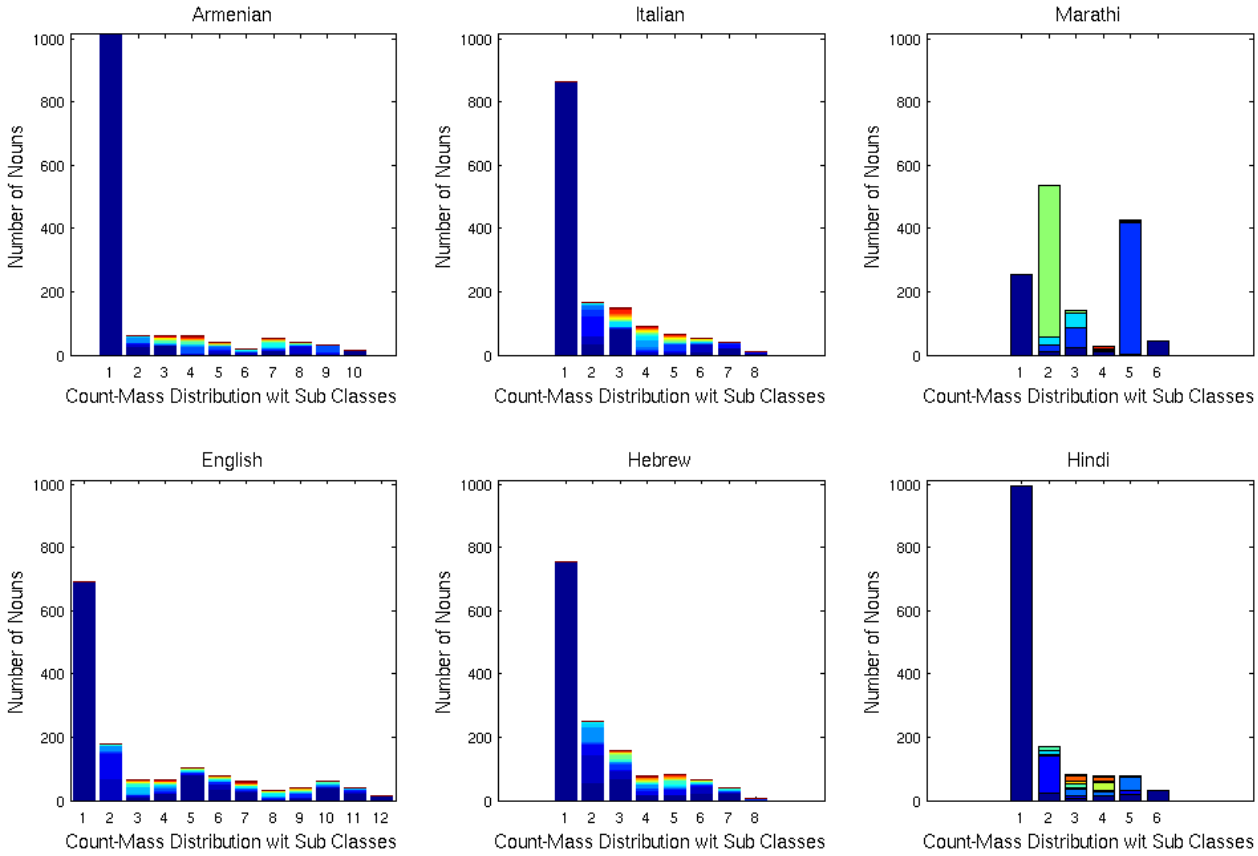

Figure 1 (Kulkarni et al 2013). Distribution of nouns along the main mass/count dimension. Each histogram reports the frequency of nouns in the database, for a particular language, at increasing distances from pure count usage (1) and towards pure mass usage (N+1), where N is the number of syntactic question for the language. Shades within the bars indicate the proportion of nouns in each of the syntactic classes that differ by the same number of features from the pure count string but do not match on which exact feature differs, hence appear collated at the same Hamming distance from the pure count.

Fig.1 shows that for 5 out of the 6 languages considered the mass count distinction is not even close to binary.It is not bimodal and it shows a substantial number of nouns in the 'pure count' class, and then a decreasing proportion of nouns at increasing distance from pure counts, but distributed among several different classes. The case of Marathi is special, and has been discussed in Kulkarni et al (2013).

2.2. Entropy and Mutual Information measures

A more detailed comparison between the languages, which preserves the multidimensional information, was obtained by measuring the mutual information between languages and the entropy of a language. Entropy quantifies the amount of variability, hencepotential 'information', contained in a system (here, on how nouns are clustered according to their usage, as defined by the binary strings), whereas mutual information quantifies how much of this information is shared between systems; for example, between a pair of languages, indicating the extent up to which clustering is similar between the pair of languages. Higher entropy means that a language has a high number of significant clusters thus pointing towards a rich classification of nounsin that language, whereas high mutual information would imply that two languages agree to a high degree on how nouns should be classified in the mass-count domain. An illustrative example is a binary variable that takes the value 1 with a probability of 0.5, implying that it has two equiprobable classes, and has the entropy of 1 bit. If instead it takes the value 1, for example, in 20% of the cases, it has the lower entropy of $\log_2(5)$-$0.8*\log_2(4) = 0.72$ bits. Thus, in a hypothetical case where the mass-count data was simply divided into two classes of mass (20% of the nouns) and count (the remaining 80%), the entropy for that dataset would be 0.72 bits. Mutual information quantifies the similarity of clustering within two different datasets, thus it will be maximal if two classifications match exactly, and it will be also 0.72 bits. Normalized to the entropy value, it will be 1, i.e., 100%. Similarly, mutual information can quantify the correspondence between syntactic classes and semantic properties. If in the example above a given semantic property could fully ‘explain’ the 20%-80% dichotomy observed in syntax, again the normalized mutual information between syntax (for a given language) and semantics would be 1.

The two main findings reported by Kulkarni et al (2013) are the high entropy values in individual languages and the low normalized mutual information values between languages, or between the mass/count syntax of any given language and semantics.

Language	Entropy
*Armenian	2.29
*Italian	3.02
*Marathi	2.71
English	3.92
Hebrew	3.40
Hindi	2.12
*Semantics	3.72 2.94(C) 2.34 (A)

Table 1 (Kulkarni et al 2013). Language–entropy relations. Entropy values in the six languages and in semantics. The * sign indicates an 'average' over five informants (three for Marathi), taken by assigning to each question and each noun the yes/no answer chosen by the majority. For semantics, the overall value (in parenthesis) has little significance, because concrete nouns areassigned to eight distinct groups and abstract to only three, and combining them distributes the abstract nouns into the two extreme concrete groups and one central group.

Table 1 shows that the Mass/Count entropy of natural languages is in the range of 2–4 bits, which indicates the presence of the entropic equivalent of 2^2–2^4 equi-populated classes of nouns (from slightly above 4 for Hindi to just below 16 for English).In practice, syntactic classes are far from being equipopulated, i.e. including each the same number of nouns, which implies that effectively there are many more than 4 classes for Hindi or 16 for English.This quantitativelyillustrates the diversity prevalent in the mass-count syntax, far beyond a dichotomous categorisation, which would have resulted in entropy values around or in fact below 1 bit.

Surprisingly, the correspondence between languages, as quantified by the normalized mutual information, was found by Kulkarni et al (2013) to be low, roughly in the range of 0.1-0.3. That is,

although all the language considered present a rich mass/count syntax, its details do not match from language to language. Correspondingly, but perhaps even more surprising from an intuitive viewpoint, also the mutual information between syntax in any language and semantics is comparably low, in the same range. This is to be expected from the low correspondence across languages, since semantic features are defined in a language-independent manner, hence they cannot simultaneously correspond to syntactic classifications that differ from each other. However, it is still striking that semantics appears to account, quantitatively, for no more than roughly 20% of the variability in syntax. This low proportion does not change much depending on how the exact measure is derived. Rather than recapitulating the full results reported by Kulkarni et al (2013), we focus here on a simple measure which can be easily analysed: the mutual information between a single syntactic marker and a single semantic feature.

Language	++	+–	–+	—	H(Lang)	H(Sem)	MI(S,L)	Norm MI
Armenian	24	31	686	43	0.451	0.366	0.080	0.218
Italian	26	29	662	67	0.536	0.366	0.053	0.145
Marathi	25	30	559	170	0.819	0.366	0.020	0.054
English	29	26	668	61	0.503	0.366	0.046	0.126
Hebrew	29	26	682	47	0.447	0.366	0.055	0.150
Hindi	28	27	686	43	0.434	0.366	0.062	0.170

Table 2. An example of the low correspondence between a semantic attribute and a syntactic property.

Semantic question 8, applied only to 784 concrete nouns, asked whether the noun denotes an entity (or individual quantity) that can be mixed withitself without changing properties. This somewhat loosely phrased question makes reference to homogeneity and cumulativity properties, since it can be interpreted either as asking whether proper parts can be permuted without changing the nature of the object, or whether instantiations can be collected under the same description. The syntactic

question considered concerned the most fundamental syntactic property of count nouns:whether the noun can be used with numerals, and it was present in all languages. The largest group of concrete nouns, in the –+ class, denote objects that are not homogeneous, and the nouns can be used with numerals. The relative proportion of nouns in each of the four classes, however, yields meagre normalized information values, indicating that individual attributes are insufficient to inform correct usage of specific rules.

In general, therefore, while the graded distribution is similar across languages, syntactic classes do not map onto each other, nor do they reflect, beyond weak correlations, semantic attributes of the concepts, as quantified by the low values of Mutual Information on both the MC dimension as well as the total mutual information. These findings are in line with the hypothesis that much of the mass/count syntax emerges from language-specific processes, or language-specific decisions as to what features of objects are relevant for realising a predicate as grammatically count.,

3. Network modelling

The goalof the second study (Kulkarni et al, 2016) was to assess the learnability of syntactic and semantic features of the mass-count distinction using simple neural networks. Artificial neural networks have a long history as a method for neurally plausible cognitive modelling (Elman 1991, Nyamapfene 2009), and can be endowedwith properties including feature extraction, memory storage, pattern recognition, generalisation, fault tolerance. Understanding how humans might acquire the capacity for handling syntax in a specific sub-domain might start from encoding syntactic/semantic knowledge into a neural network, which self-organizes with a prescribed learning algorithm to recode that information in a neurally plausible format. That way one may draw parallels about governing principles in the brain that bring about the acquisition of syntax. Taking cues from biological neurons, most artificial neural networks employ ‘Hebbian’ plasticity rules, wherein the synaptic connection between two units is strengthened if they are activated nearly simultaneously,

thus leading to associative learning of the conjunction or sequence of activations.

We have considered a competitive network, a simple self-organising network which through ‘unsupervised’ learning may produce a useful form of recoding. A competitive network, under the right conditions, is able to discover patterns and clusters in a stimulus space and to train itself to correctly identify and group inputs that share a close resemblance to each other. A competitive network is particularly interesting in our case since much linguistic information during language acquisition is 'discovered' rather than explicitly taught. Moreover, mass and count nouns have been shown to exhibit differential evoked potential responses, both with a syntactic and with a semantic stimulus (Chiarelli et al 2011). The performance of a simple competitive network should indicate how well syntactic and semantic features can be accommodated within a single network, thus exploring if the network can indeed achieve some pattern recognition that will allow it to successfully categorise nouns in the syntactic mass-count space.

3.1. Methods

Our study is reported in detail in (Kulkarni et al, 2016) and here we just recapitulate the key ideas and results. The network consists of a single input and a single output layer. At the input layer each unit represents a syntactic feature (‘numeral’, ‘a/an’ etc) in case of the syntactic network or a semantic feature (‘fixed shape’, ‘fluidity’ etc) for the semantic network. The input layer is binary, and for each noun given as input a given unit can be active (activation value 1) to indicate that the feature can be attributed to the noun, or inactive (value 0) to indicate that it cannot. Thus a single learning event for the network includes the application of a binary input string containing the syntactic or semantic information pertaining to a single noun, activity propagation to the output units, and modification of the synaptic weights according to the prescribed learning rule. In a variant to be considered later, instead of self-organizing an output representation of nouns, we explore the self-organization of syntactic features (‘markers’); in that variant, rather than an input nounwith the features as components, we apply as input a single feature/marker, with the nouns as components, i.e. there are

a few very long input string instead of many short strings. On the output side, the number of units is variable, set by the simulation requirements. Unlike the input units, outputs units are graded, taking continuous values in the range of 0 to 1. A competition amongst the output units based upon their activation levels decides the final output level of each unit.

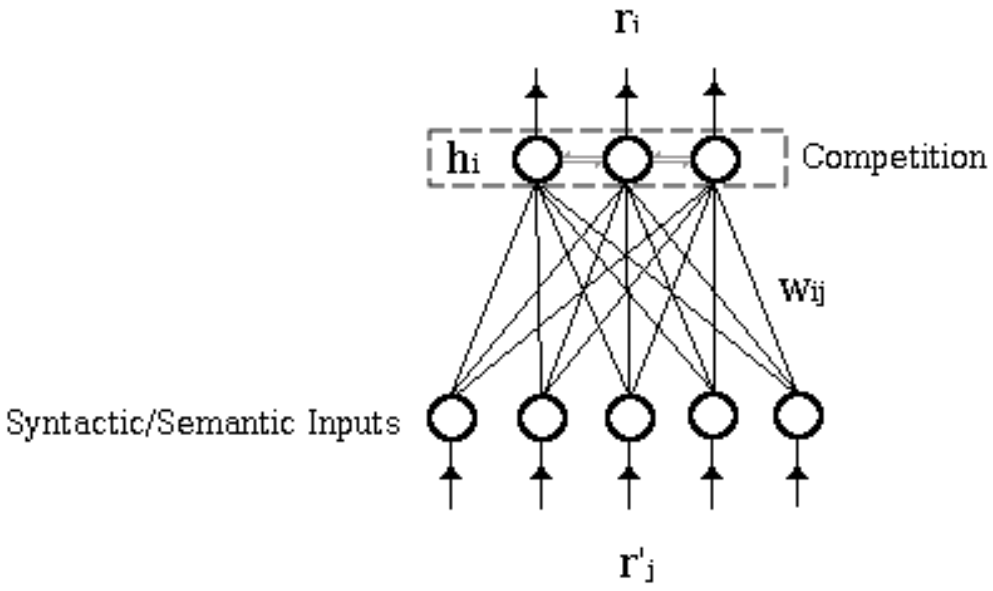

Figure 2. Schematic diagram of the artificial neural network, showing an input layer where units are binary strings containing syntactic/semantic information of nouns and an output layer where units compete with each other to produce graded firing rates based on connection weights and on competition.

The weights in the network are adjusted during a learning phase according to a so-called Hebbian rule, taking into account the input and output firing rates of the units connected by that weight. The learning rule is slightly modified from the standard Hebbian rule to incorporate normalisation of the weights during learning in a biologically plausible way.

One training iteration includes presenting each noun in the list once and the above process is repeated for the desired number of iterations. A softmax function is implemented at the output stage to facilitate competition amongst output neurons which consequently lead to the strengthening of the connections

between output neurons with the maximum firing rate and relevant input neurons.

We again use mutual information as a measure to analyse the correspondence between two representations, encoded either in a syntactic network trained on input information about marker usage for the nouns in a particular language, or in a semantic network trained on information about the semantic properties of the nouns (Kulkarni et al 2013). Here we focus on systems that have undergone a slow process of self-organisation to categorise their inputs.

3.2. Results: Classification of Markers

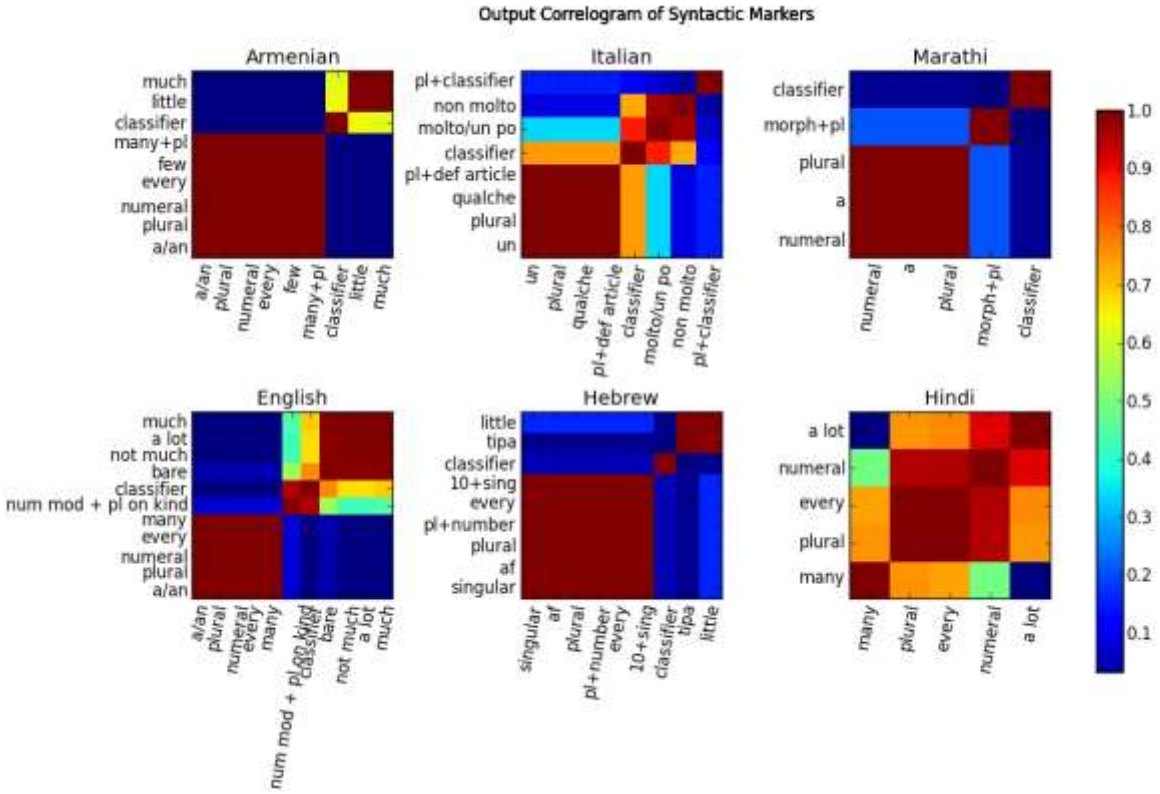

Figure 3. Correlograms for 784 concrete nouns in each of the 6 languages in our study. Dark blue regions represent complete lack of correlation (orthogonal vectors) while dark red regions represent congruent vectors. Markers that are syntactically identified as "Count" tend to be highly correlated at the output of the network. To an extent, the result is similar with markers that may be syntactically identified as "Mass".

First, in what we earlier called a variant of the standard network approach, we present as input the syntactic markers used in the classification of the nouns. Here an input vector is comprised of *n* units, where*n* is the number of nouns (784 in case of concrete nouns and 650 in case of abstract nouns), for each of the syntactic markers. Thus an input includes information on how that particular marker is used over all the nouns. Each input vector is presented once in one iteration, for 50 such iterations, which is also when the synaptic weight matrix is observed not to change with further iterations. After obtaining the output firing rates for each input marker at the end of the iterations, we calculate the correlogram, representing how correlated the output vectors are with each other, hence giving information about marker categorization. We show the mean correlograms over 50 distinct network simulations.

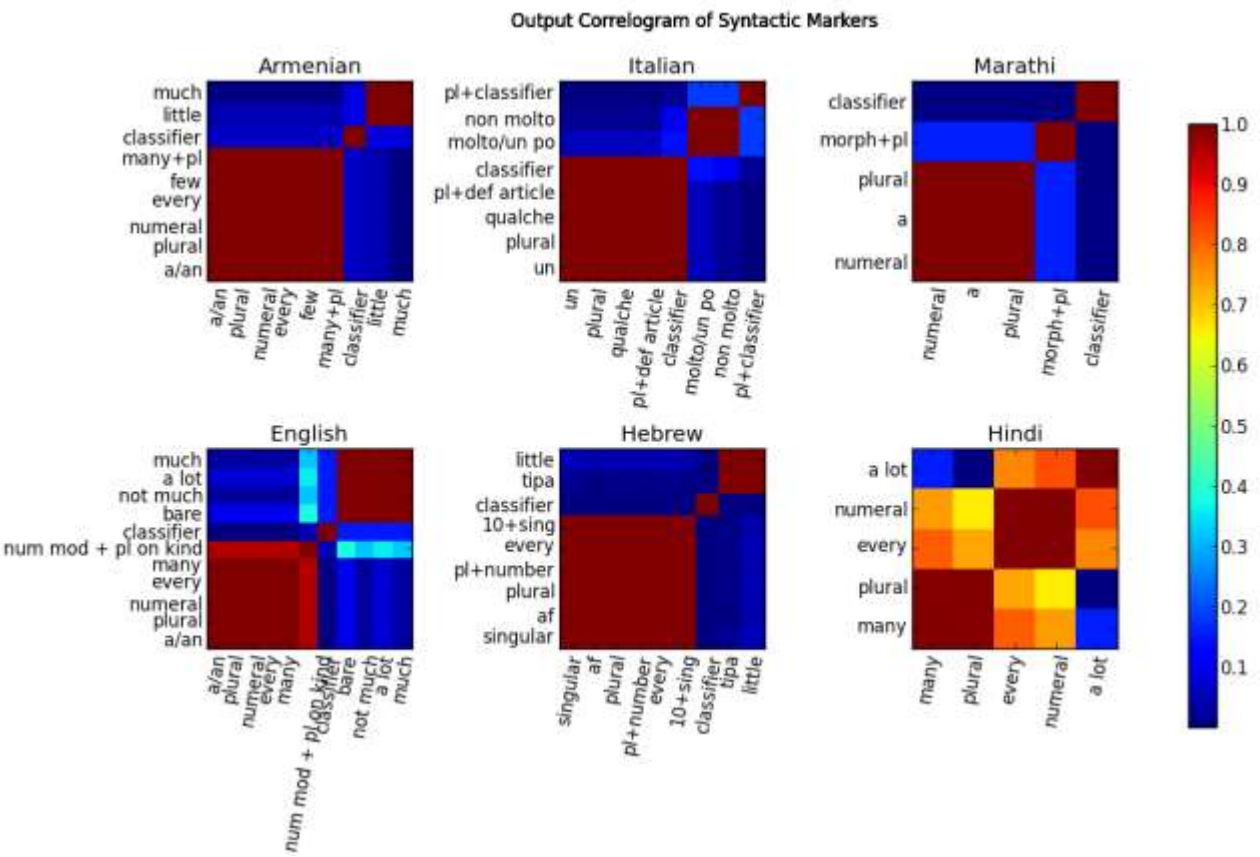

Figure 4. Correlograms, same as in Figure 3, but for 650 abstract nouns. Note that markers are ordered in the same way as in Fig.3.

The correlograms in Fig.3 allow to visually identify markers that fall in the same category, as self-organized in the output of the network. High levels of correlation between two markers signify close proximity in the firing rates of the output units for that pair of markers, and are represented by warmshades towards brown. For concrete nouns in Armenian, markers like ‘a/an’, ‘plural’, ‘numeral’, ‘few’, ‘every’ and ‘many+plural’ have a correlation of 1, thus occupying the same position in the output space of the network. These are markers that can be applied to count nouns and not to mass nouns. Instead, the typical mass markers of ‘measure classifier’ form an independent representation, whereas ‘little’ and ‘much’ share the same position in output space but distant from the count markers. Italian, Marathi, English and Hebrew follow the same Armenian line of grouping count markers together and having separate but nearby representation for mass markers, distant from the count markers. Hindi is different, as 4 of its 5 chosen markers appear to be ‘count’ in nature, but all show gradation within the broad count category. Results are similar for abstract nouns except for Italian

having fewer graded categorisation than for concrete nouns (Fig.4).

The competitive network can be similarly tested on semantic features based on what value each feature assumes over all the nouns. As seen in figure 5, semantic features are neatly divided into mass and count features. Count features like ‘single unit’, ‘boundary’, ‘stable shape’ and ‘degradation’ all have a correlation of 1 with each other and 0 with mass features like ‘free flow’, ‘container shape’ and ‘mixing’. While ‘free flow’ forms a separate representation, ‘container shape’ and ‘mixing’ have the same output activation.

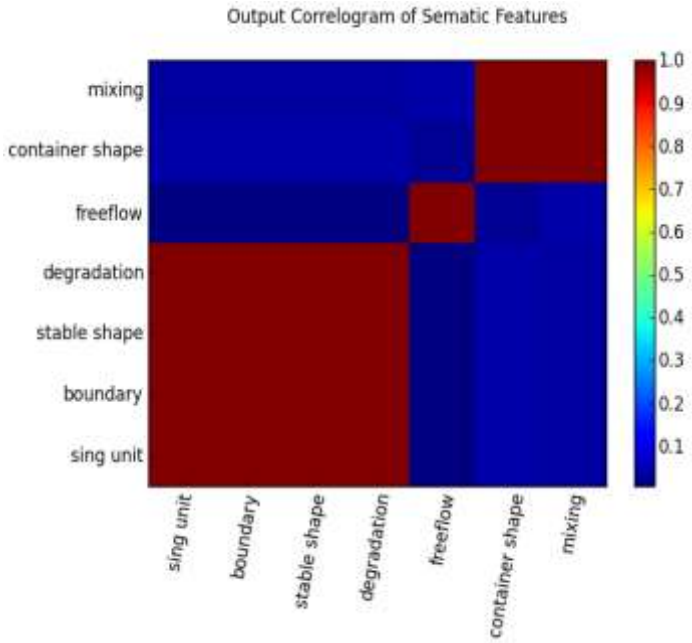

Figure 5. Correlogram of semantic markers for concrete nouns. Physical properties associated with “countable” nature of nouns are clearly separated from properties that can be associated with “fluid-like” properties.

3.3. Results: Categorization of nouns

Similar to the procedure in section 3.2, we now present nouns as input to the network and analyze the activation of the output units. The input vector here consists of *n* units for each noun, i.e., the number of markers for a language, hence containing information on how the noun is used over all the mass-count markers for that language. Figure 6 shows the position of the nouns in the 3-D output space, where each axis represents an output unit. Axes are selected such that x, y and z, respectively,

represent units in descending order of variance over the values of output activation they span. The shadeof each point signifies where that noun (or cluster of nouns, since nouns classified as identical are co-incident) lies on the MC dimension as defined by the Hamming distance from the pure count string (see section 1.1 A). Blackindicates a distance of 0, thus pure count, while whiteindicates a distance of 1, representing a ‘mass noun’.

Nouns are seen to approximately fall along a single line for all languages (a predominantly linear structure for English), barring an outlier at 0 which represents inputs, all of which are inactive for a noun. Moreover we can see a gradient from blackto white, which implies that nouns, even though not completely faithful, to a great extent lie along a gradient from ‘count’ to ‘mass’. We further visualise the distribution of nouns on this line, so as to assess the frequency of nouns in each cluster. The axis with maximum variance is selected and a histogram of the number of nouns in each cluster along this axis is plotted.

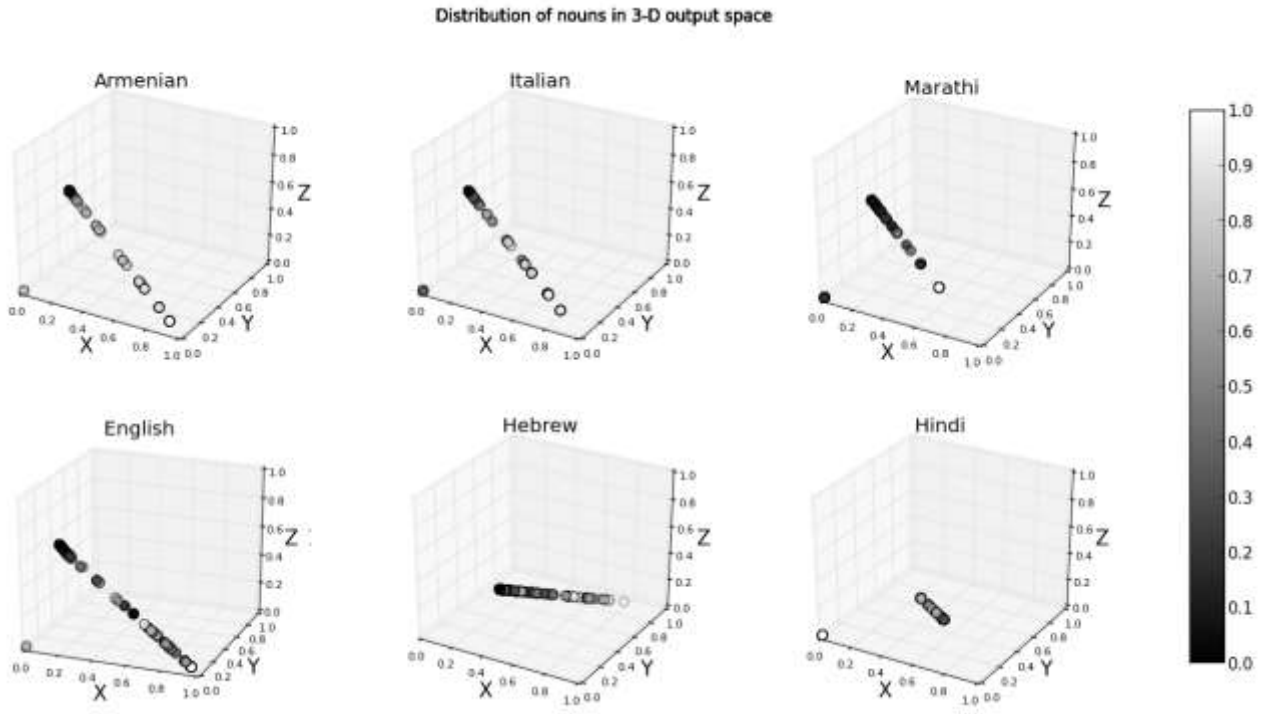

Figure 6. Position of 784 concrete nouns in the output space as defined by 3 output units in 6 languages. The gray scaleindicates the Hamming distance of the noun on the MC dimension, from black = ‘pure count’ to white = ‘pure mass’.

Is it interesting to note that a dimensionally reduced, entropy preserving representation of the mass-count nouns has a notional similarity to the concept of the MC dimension as in section 1.1A, Figure 1. The MC dimension was introduced as a concept to better understand the mass-count division in terms of the 'pure count' string, but a competitive network with the appropriate parameters is able to bring about a roughly similar distribution without needing a prior definition ad hoc.

Results for abstract nouns are similar (see Kulkarni et al, 2016).

3.4. The syntax-semantics interaction

As we saw from the information theoretical analysis, the syntax and semantics of the mass-count distinction share only a weak direct link, in the core structure of the count class (Kulkarni et al 2013), at least based on the semantic properties which we chose to investigate (see Appendix). Thus acquiring the complete set of syntactic classes from semantic classes is not possible by any learning mechanism, due to a lack of a direct one-to-one correspondence. However it is improbable that syntax and semantics are independently learned without any mutual interaction during learning of mass-count concepts, and there is no evidence that either one is learnt before the other (Nicolas 2010). From the classification of markers above, we see that broad categories of mass and count can indeed be extracted out of the data, interestingly for both syntax and semantics, thus rendering some semantic sense to the syntactic distinction. Classes of nouns formed from these markers do not reflect, however, mass-count information in a straightforward manner between syntax and semantics. Hypothesising an underlying commonality of the mass-count divide between markers of syntax and semantics,we thustested the performance of the competitive network when syntax and semantics are simultaneously part of the input space duringthe learning phase, and test the correspondence between the syntactic and semantic classes after learning.

First, to compare with our previous results, we calculated the baseline mutual information between syntax and semantics by providing only semantic information to the network, with no syntactic information during the learning phase. The mutual information was calculated between syntactic data

and the output of the semantic competitive network. When no syntactic information is present at the inputs, the resulting mutual information is about equal to the mutual information between the syntactic and semantic data, as calculated using the procedure in Kulkarni et al (2013).

The competitive network brings about a dimensional reduction from a high dimensional input space to a lower dimensional output space defined by the number of output units.

Syntactic information was then provided to the network in a partial manner, in a proportionγ, which signifies the fraction of input units of the syntactic segment, of the input string, that are set to the activation levels of the syntactic string of a particular language. γ=0 corresponds to when none of the syntactic input units are receiving any information and are set to 0; while γ=1 implies that all of the syntactic information is present; for in-between cases a fraction 1-γ units are randomly selected and set to 0. Thus we were able to vary the amount of syntactic information available to the network during learning and test the effect on the syntactic-semantic mutual information and whether the relevant syntactic and semantic classes are brought together in any systematic way. We trained and tested the network by providing the same proportion γ of syntactic inputs along with the semantic ones.

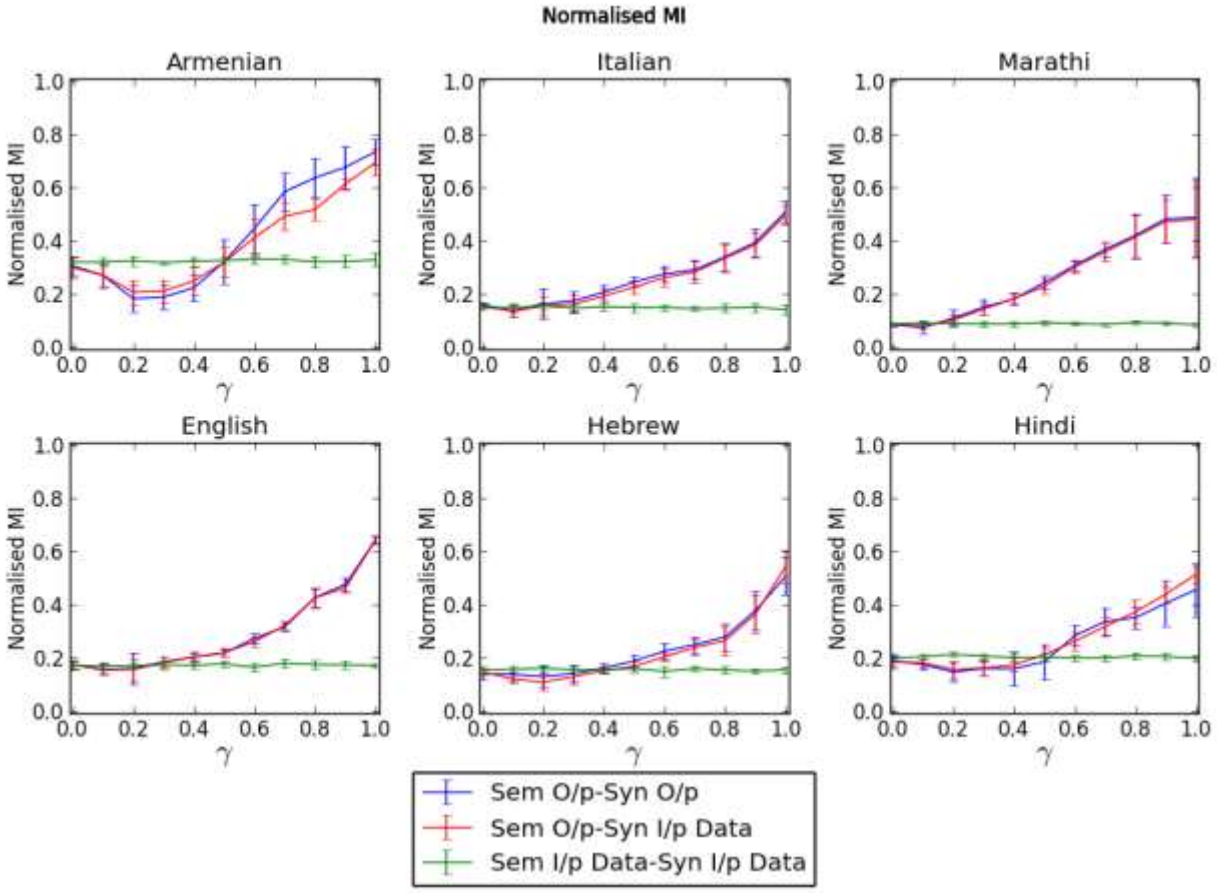

Figure 7. Mean normalised mutual information over 10 independent simulations at various γ for concrete nouns in 6 languages when γ=1, N =3, 10 iterations.

The results were disappointing from the point of view that the combined syntactic-semantic inputs have only a limited influence on learning. Figure 7 depicts the performance of the network when it is tested on the semantic inputs while they are incrementally supplied with syntactic information. The green curves represent the mutual information between unprocessed semantic and syntactic input, this is the baseline mutual information between semantic and syntactic data which remains flat, i.e. independent of γ. Unprocessed inputs consist of 'binary strings',whichare the raw data. The small fluctuations seen in the baseline curve are a result of variation produced due to the sampling correction in calculating mutual information for each independent simulation (Kulkarni et al 2013). The red curves represent the mutual information between the self-organised output of the network and unprocessed syntactic inputs while the blue curves plot the mutual information between self-organised semantic and syntactic outputs. The red and blue curvestend to follow each other closely,

implying that the syntactic competitive network results in a dimensionally reduced faithful representation of the syntactic input data. The mutual information rises above the baseline as γ increases above the 0.4-0.5 region for Armenian, English, Hebrew and Hindi, and above the 0.2 region for Italian and Marathi.Thus semantic classes tend to gradually realign, with increasing γ, in such a manner that they correspond more to the syntactic classes as compared to the baseline. This reorganisation is however very limited: at γ=1, the normalised mutual information for all languages is in the range of 0.5-0.6, which is around half way towards full agreement. Although interacting with syntax does help some reorganisation of the semantic classes, the divide between syntax and semantics is clear and almost half of the semantic information cannot be shared with syntax at γ=1..

The performance of the network is further limited by the fact that it cannot be driven by semantic units only, with nosyntactic information during testing. A 'syntactic context' is necessary at the inputs for the network to result in a mutual information performance above baseline. When tested without syntactic information, or with only a partial amount, the drop in the normalised mutual information is significant, with only a tiny trace of learning shown by the network.

Thus, as a summary, we can see that a basic unsupervised neural network implementing a Hebbian-like learning rule and soft competition amongst output neurons is: a) successful in classifying the syntactic and semantic markers into two separate classes (in most cases), of "mass" and "count" – implying that a feature, based on how it is used by native speakers over various nouns can represent the notion of countability or uncountability; b) partly successful in discovering a single "Mass-Count" dimension in the data and the spectral nature of the nouns along this dimension through the usage of nouns across different syntactic features; c) unsuccessful in its ability to predict syntax from semantics (or vice-versa), hence not discovering a map across these two domains.

4. Conclusions

This is clearly a first attempt at exploring the learnability of this specific sub-domain with a simple neural network. The results lead to a number of inferences.

1. In most languages, syntactic markers tend to categorize ‘spontaneously’ between mass and count markers, lending validity to the intuitive perception of a quasi-binary distinction. This is not fully true, however, and particularly in Hindi the markers chosen show a graded distribution of mutual correlations.

2. Nouns, instead, tend in most languages to distribute quite closely along a line which coincides with the main mass/count dimension introduced in our previous study (Kulkarni et al 2013). Along this line, nouns are very crowded at the count end, and scattered all along towards the mass end. Their distribution is therefore graded rather than binary, with no emergence of a single ‘mass’ class, but rather of several non-exclusive but distinct ways for a noun to be different from pure count. For example, in Armenian (Figure 8) nouns like ‘bird’ and ‘ship’ belong to the ‘pure count’ class while ‘troop ’and ‘lunch’ are in the 9th class away from the count class. On the mass end, nouns like ‘cotton’ and ‘milk’ are at the extreme mass end of the spectrum while ‘coffee’ and ‘wheat’ are more mass-like nouns but not at the pure mass end. The exception is English, where there are at least two clear non-equivalent dimensions of non-countability.

Both the above observations are interesting because the mass-count information in thecategorisation arises on its own. The markers, in some cases, very cleanly segregate themselves intomass and count. The nouns are reduced to a one dimensional representation along a mass-countspectrum. Even though the network fails to associate specific syntactic markers with specific nounsbased on the semantics, the network does develop a 'concept', if we may say, of what the mass-countclassification is. The diversity and richness of this classification however, prevents a simplenetwork to learn specific associations. This brings us to the third observation,

3. Finally, the lack of significant mutual information between semantics and syntax implies, as

we have verified, that the latter cannot be extracted solely from the former. Further, when allowing the competitive network to self-organize on the basis of full semantics and partial syntactic inputs, and testing it with the full syntactic inputs, the mutual information obtained with the full syntactic usage distribution is only at most about half the corresponding entropy value. This occurs in fact only when the full syntax is given in the input also at training, and it indicates that giving also semantics information affects negatively rather than positively the performance of the network.

Overall, these observations do not clarify how mass count syntax may be acquired by humans with neurally plausible mechanisms, on the basis of matching semantic information and syntactic properties. They reinforce the conclusions of our earlier study, that mass count syntax is far from a rigid binary contrast. It appears as the flexible, language-specific and even, when within-language,speaker-specificusage of a variety of binary markers to a quantitatively and qualitatively graded repertoire of nouns, where being non-count can be expressed in many ways. Furthermore, the count-mass contrast cannot be derived from a set of general, abstract semantic features, such as those listed in Appendix A. One possibility, suggested by the high clustering of nouns in individual languages, but lack of mutual information between languages, is that the semantic features we chose to work with are supplemented by a series of more specific, less abstract features, salient in a particular language.In general, our results support recent work by Rothstein (in press) arguing that the count/mass contrast is not a reflection of a contrast between atomicity and homogeneityor between objects and stuff. Instead, it reflects a perspectival contrast between entities presented grammatically as countable and those presented as contextually non-countable.Taking the abstract properties of the referents into account is of limited use in generating grammatical generalizations.

REFERENCES

Bale A.C. and Barner D. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. Journal of Semantics 26,217–252.

Chiarelli V., El Yagoubi R., Mondini S., Bisiacchi P. and Semenza C., 2011. The syntactic and semantic processing of mass and count nouns: An ERP study. PloS one, *6*(10), e25885.

Chierchia G. 2010. Mass nouns, vagueness and semantic variation. Synthese 174, 99–149.

Elman J. L. 1991.Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning, 7:195-224.

Kulkarni R., Rothstein S. and Treves A. 2013. A statistical investigation into the cross-linguistic distribution of mass and count nouns: Morphosyntactic and semantic perspectives. Biolinguistics, 7, pp.132-168.

Kulkarni R., Rothstein S. and Treves A. 2016. A Neural Network Perspective on the Syntactic-Semantic Association between Mass and Count Nouns. Journal of Advances in Linguistics, Vol 6, No 2

Link G. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In R.B. Bäuerle, C. Schwarze and A. von Stechow (eds.) Meaning, Use and Interpretation, 303–323. Berlin: Mouton de Gruyter. [Reprinted in P. Portner and B. Partee (eds.). 2002. Formal Semantics: The Essential Readings, 127–146. Oxford: Blackwell.]

Nicolas D.A. 2010. Towards a semantics for mass expression derived from gradable expressions.

Recherches Linguistiques de Vincennes 39, 163–198

Nyamapfene A. 2009. Computational investigation of early child language acquisition using multimodal neural networks: a review of three models. Artificial Intelligence Review 31 (1-4): 35-44.

Pelletier F.J. 2010. Descriptive metaphysics, natural language metaphysics, Sapir-Whorf, and all that stuff: Evidence from the mass-count distinction. Baltic International Yearbook of Cognition, Logic and Communication, 6(1), p.7.

Pinker S. 1995. The language instinct: The new science of language and mind (Vol. 7529). Penguin UK.

Prasada S., Krag F. and Haskell T. 2002. Conceiving of entities as objects and stuff. Cognition 83, 141–165.

Rothstein S. 2017. Semantics for counting and measuring. Cambridge University Press.

Soja N.N., Carey S. and Spelke E.S. 1991. Ontological categories guide young children's inductions of word meanings: Object termsand substance terms. Cognition 38, 179–211.

Appendix A: Semantic Questions

1. Is it Alive irrespective of context?

2. It is an Abstract Noun?

3.Does it have a single Unit to represent itself?

4.Does it have a definite Boundary, visually or temporally?

5.Does it have a stable Stationary shape (only if concrete)?

6.Can it Flow freely (only if concrete)?

7.Does it take the shape of a Container (only if concrete)?

8.Can it be Mixed together indistinguishably (only if concrete)?

9. Is the identity Degraded when a single unit is Divided (only if concrete)?

10. Can it have an easily defined Temporal Unit (only if abstract)?

11.Is it an Emotion /Mental process (only if abstract)?

12. Can it have an easily defined Conceptual Unit (only if abstract)?